Matrix Calculus

CSCI 5521 Machine Learning fundamentals

Some Definitions

- Quadratic Form
  - A: a nxn square matrix ∈ $R^{n \times n}$
  - x: a nx1 vector ∈ R^n
  - the ***quadratic form***: x^TAx
    - It is a scalar value.
    - We often implicitly assume that A is symmetric since $x^TAx = x^T(A/2+A^T/2)x$
    - If we write it as the elements of x and A, it is

$$x^T Ax = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j$$

Some Definitions

- Quadratic Form
  - example

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$x^T A x = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ax_1 + cx_2 & bx_1 + dx_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = ax_1^2 + bx_1x_2 + cx_1x_2 + dx_2^2$$

Some Definitions

- Positive Definite (PD)
  - A: A symmetric matrix $\in S^n$
    - For all **non-zero** vectors $x \in R^n$, $x^TAx > 0$.
    - Then A is ***positive definite*** (PD)
- Positive Semidefinite (PSD)
  - A: A symmetric matrix $\in S^n$
    - For all vectors $x \in R^n$, $x^TAx \geq 0$.
    - Then A is ***positive semidefinite*** (PSD)
- Negative Definite and Negative Semidefinite
- Indefinite

Some Definitions

- Positive Definite (PD)
  - example

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

$$x^T A x = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^3 = x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 \geq 0$$

Some Definitions

- Eigenvalues and Eigenvectors
  - A: a square matrix $\in R^{n \times n}$
  - λ: $\in C$
  - x: a vector $\in C^n$
    - If $Ax = \lambda x$, $x \neq 0$, λ is an ***eigenvalue*** of A and x is the corresponding ***eigenvector***.
    - λ is a solution to $|(\lambda I - A)| = 0$.
    - The corresponding eigenvector of λ_i is the solution to the linear equation $(\lambda_i I - A)x = 0$.
    - There are more efficient methods in practice to numerically compute the eigenvalues and eigenvectors.

Some Definitions

- Eigenvalues and Eigenvectors
  - example

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$|\lambda I - A| = \begin{vmatrix} \lambda - 2 & -1 \\ -1 & \lambda - 2 \end{vmatrix} = \lambda^2 - 4\lambda + 3 = 0 \Rightarrow \lambda_1 = 1, \lambda_2 = 3$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 3\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Properties of Eigenvalues and Eigenvectors

- The trace of a A is equal to the sum of its eigenvalues.
- The determinant of A is equal to the product of its eigenvalues.
- The rank of A is equal to the number of non-zero eigenvalues of A.
- If A is non-singular then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector x_i.
- The eigenvalues of a diagonal matrix $D = diag(d_1, \ldots d_n)$ are just the diagonal entries $d_1, \ldots d_n$.
- Diagonalizable:
  - We can write all the eigenvector equations together as $AX = X\Lambda$.
  - If the eigenvectors of A are linearly independent, $A = X\Lambda X^{-1}$. We say A is ***diagonalizable***.

Eigenvalues and Eigenvectors of Symmetric Matrices

- A: a symmetric matrix $\in S^n$
  - All the eigenvalues of A are real.
  - The eigenvectors of A are orthonormal (The inner product is 0.).
  - A is diagonalizable: $A = U\Lambda U^T$ (Note: $U^{-1} = U^T$)
    - $x^T A x = x^T U \Lambda U^T x = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2$
    - All $\lambda_i > 0 \Rightarrow$ A is positive definite
    - All $\lambda_i \geq 0 \Rightarrow$ A is positive semidefinite
    - A has both positive and negative eigenvalues $\Rightarrow$ A is indefinite

What is Matrix Calculus

- Calculus
  - Differential calculus
    - Derivative
      - e.g. $f(x)=x^2$, derivative function $f'(x)=2x$
  - Integral calculus
- Matrix Calculus
  - Extension of calculus to the vector/matrix setting
    - Gradient
    - Hessian

The Gradient

- Definition
  - Function f : $R^{m \times n} \rightarrow R$
  - A: m × n matrix
  - The ***gradient*** of f (written as $\nabla_A f(A)$) is an m × n matrix and each element of the matrix is a partial derivative defined by

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}$$

The Gradient

- Example
  - A: 2x2 matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$
  - f(A)=|A|
  - calculate each element of $\nabla_A f(A)$

$$(\nabla_A f(A))_{11} = \frac{\partial f(A)}{\partial A_{11}} = \frac{\partial(a_{11}a_{22} - a_{12}a_{21})}{\partial a_{11}} = a_{22}$$

$$(\nabla_A f(A))_{12} = \frac{\partial f(A)}{\partial A_{12}} = \frac{\partial(a_{11}a_{22} - a_{12}a_{21})}{\partial a_{12}} = -a_{21}$$

$$(\nabla_A f(A))_{21} = \frac{\partial f(A)}{\partial A_{21}} = \frac{\partial(a_{11}a_{22} - a_{12}a_{21})}{\partial a_{21}} = -a_{12}$$

$$(\nabla_A f(A))_{22} = \frac{\partial f(A)}{\partial A_{22}} = \frac{\partial(a_{11}a_{22} - a_{12}a_{21})}{\partial a_{22}} = a_{11}$$

The Gradient

- Example
  - The gradient of f

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} \end{bmatrix} = \begin{pmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{pmatrix}$$

  - The general case for f(A)=|A|

$$\nabla_A f(A) = |A| A^{-T}$$

The Gradient

- When A is a vector
  - a vector $x \in R^n$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

  - the gradient of f

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

- Two properties
  - $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$
  - For $t \in R$, $\nabla_x (t\, f(x)) = t\nabla_x f(x)$
- Two important notes
  - $\nabla_A f(A)$ is always the **same** as the **size** of A
  - the gradient of f is defined only if f is a real-valued function
    - e.g. we can't take the gradient of f=2A with respect to A

The Hessian

- Definition
  - Function f: $R^n \rightarrow R$
  - x: an nx1 vector
  - The ***Hessian*** matrix with respect to x (written as $\nabla_x^2 f(x)$) is an n × n matrix and each element of the matrix is a partial derivative defined by

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

The Hessian

- Example
  - x: a 2x1 vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$
  - f(x)= $x^T \begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix} x$
  - calculate each element of $\nabla_x^2 f(x)$

$$(\nabla_x^2 f(x))_{11} = \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} = \frac{\partial^2 (x_1^2 - x_1 x_2 + 4x_2^2)}{\partial x_1 \partial x_1} = \frac{\partial(2x_1 - x_2)}{\partial x_1} = 2$$

$$(\nabla_x^2 f(x))_{12} = \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} = \frac{\partial^2 (x_1^2 - x_1 x_2 + 4x_2^2)}{\partial x_1 \partial x_2} = \frac{\partial(2x_1 - x_2)}{\partial x_2} = -1$$

$$(\nabla_x^2 f(x))_{21} = \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} = \frac{\partial^2 (x_1^2 - x_1 x_2 + 4x_2^2)}{\partial x_2 \partial x_1} = \frac{\partial(-x_1 + 8x_2)}{\partial x_1} = -1$$

$$(\nabla_x^2 f(x))_{22} = \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} = \frac{\partial^2 (x_1^2 - x_1 x_2 + 4x_2^2)}{\partial x_2 \partial x_2} = \frac{\partial(8x_2)}{\partial x_2} = 8$$

The Hessian

- Example
  - The Hessian matrix of f

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 8 \end{pmatrix}$$

- In general, if $f(x) = x^T A x$ and $A \in S^n$,

$$\nabla_x^2 f(x) = 2A$$

The Hessian

- Some notes
  - The Hessian is defined only when f(x) is real-valued.
  - Hessian is always **symmetric**.
  - We will only consider taking the Hessian with respect to a vector.
  - The Hessian is **not** the gradient of the gradient.
    - However, the gradient of the ith entry of $\nabla_x f(x)$ is the ith column (or row) of $\nabla_x^2 f(x)$.
- Some useful results
  - $\nabla_x b^T x = b$
  - $\nabla_x x^T A x = 2Ax$ (if A symmetric)
  - $\nabla_x^2 x^T A x = 2A$ (if A symmetric)

Application in Least Squares Optimization

- The problem
  - Given a full-ranked matrix $A \in R^{m \times n}$ and a vector $b \in R^m$
  - Suppose there is no x such that Ax=b.
  - Find a vector $x \in R^n$, such that the square of the Euclidean norm $||Ax - b||_2^2$ is minimized.
- Solve the problem

$$\|Ax-b\|_2^2 = (Ax-b)^T(Ax-b) = x^T A^T Ax - 2b^T Ax + b^T b$$

  - Take the gradient with respect to x

$$\nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) = \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b = 2A^T Ax - 2A^T b$$

  - Set the gradient to zero (vector) and we get the solution

$$x = (A^T A)^{-1} A^T b$$

End!